Computational Techniques for Machine Learning Assignment 2

Constanza Marini[1] and Diana Laura Aguilar[1]

Tecnologico de Monterrey, Carretera al Lago de Guadalupe Km. 3.5, Atizapán de Zaragoza, Estado de México 52926, México.
A01332485@itesm.mx
A01751168@itesm.mx

1 Assignment Description

In this assignment, the Validity Index using supervised Classifiers (VIC) [1] was used in order to evaluate the effect of the partition size in binary and tertiary partitioned datasets with a 10 cross-validation procedure. As to the VIC implementation, the python source code provided by the authors of [1] was leveraged throughout this project. The dataset used here was generated with the information of 247 fingerprints (5241 objects); each fingerprint was associated to a set of minutiae with 132 attributes, along with the score-change values. The latter attribute was selected as the target in order to make partitions according to a specific cutoff value. Moreover, the analysis was performed on 50 different partitions of 2 sets and 50 different partitions of 3 sets; the leveraged supervised classifiers were Random Forest, Decision Tree, Extra Tree, Naïve Bayes, k-NN, Multi-layer Perceptron (MLP) and Linear Discriminant Analysis (LDA). For this purpose, we utilized the implementations of scikit learn [2] with their default parameters.

2 Results and Discussion

2.1 Binary Partition

As mentioned before, the dataset was divided into two clusters according to the score-change and different cutoff values: we set 50 evenly spaced values, calculated over the interval $[-0.2, 0.2]$. In Table 2.1, the cutoff value and cluster sizes, along with the VIC value and the best supervised classifier are presented for each partition. Additionally, Fig. 1 shows the former values and Fig. 2 displays the Area Under the Receiver Operating Characteristic Curve (AUC) [3] of each classifier over each partition.

As it can be seen in Fig. 2, the class balance, or cluster size, affects the AUC value differently depending on the supervised classifier, for example, k-NN has a close to constant trend (purple line), suggesting an independence over the cluster size; while Naïve Bayes changes considerably according to the cluster size (yellow line). An interesting fact is that the classifiers with highest AUC (see Fig.

3) and hence highest VIC score were obtained in the partitions with extremely imbalanced classes. The highest VIC score (0.714) was obtained in partitions 0 and 1, with a cutoff point of -0.2 and -0.192, respectively, and a cluster size of ∼4000 vs ∼1200 observations (see Table 2.1). Furthermore, the best classifiers for this dataset were Random Forest, Naïve Bayes and LDA.

Table 1. Relevant information and results for the binary partitions.

Partition Number	Cutoff Value	Size of Class 0	Size of Class 1	VIC	Best Classifier
0	-0.2	4073	1168	0.717	Random Forest
1	-0.192	4020	1221	0.717	Random Forest
2	-0.184	3969	1272	0.714	Random Forest
3	-0.176	3907	1334	0.712	Random Forest
4	-0.167	3851	1390	0.71	Random Forest
5	-0.159	3787	1454	0.714	Random Forest
6	-0.151	3706	1535	0.704	Random Forest
7	-0.143	3622	1619	0.706	Random Forest
8	-0.135	3544	1697	0.7	Random Forest
9	-0.127	3473	1768	0.699	Random Forest
10	-0.118	3378	1863	0.704	Random Forest
11	-0.11	3299	1942	0.702	Random Forest
12	-0.102	3206	2035	0.699	Random Forest
13	-0.094	3126	2115	0.694	Random Forest
14	-0.086	3029	2212	0.69	Random Forest
15	-0.078	2949	2292	0.69	Random Forest
16	-0.069	2852	2389	0.684	Random Forest
17	-0.061	2765	2476	0.683	Random Forest
18	-0.053	2676	2565	0.684	Random Forest
19	-0.045	2568	2673	0.676	Random Forest
20	-0.037	2470	2771	0.673	Random Forest
21	-0.029	2374	2867	0.671	Random Forest
22	-0.02	2282	2959	0.664	Random Forest
23	-0.012	2165	3076	0.662	Random Forest
24	-0.004	2069	3172	0.656	Random Forest
25	0.004	1932	3309	0.653	Random Forest
26	0.012	1778	3463	0.634	Random Forest
27	0.02	1687	3554	0.642	Random Forest
28	0.029	1580	3661	0.63	Random Forest
29	0.037	1482	3759	0.631	Random Forest
30	0.045	1372	3869	0.64	Random Forest
31	0.053	1282	3959	0.631	Random Forest
32	0.061	1191	4050	0.632	Random Forest
33	0.069	1120	4121	0.627	Random Forest
34	0.078	1055	4186	0.628	Random Forest
35	0.086	996	4245	0.627	Random Forest
36	0.094	938	4303	0.626	Random Forest
37	0.102	885	4356	0.638	Random Forest
38	0.11	831	4410	0.624	Random Forest
39	0.118	773	4468	0.625	Naive Bayes
40	0.127	706	4535	0.634	Naive Bayes
41	0.135	659	4582	0.639	Naive Bayes
42	0.143	619	4622	0.647	Naive Bayes
43	0.151	573	4668	0.649	Naive Bayes
44	0.159	528	4713	0.651	Naive Bayes
45	0.167	501	4740	0.662	Naive Bayes
46	0.176	476	4765	0.671	Naive Bayes
47	0.184	435	4806	0.689	Random Forest
48	0.192	413	4828	0.682	Naive Bayes
49	0.2	386	4855	0.684	Naive Bayes

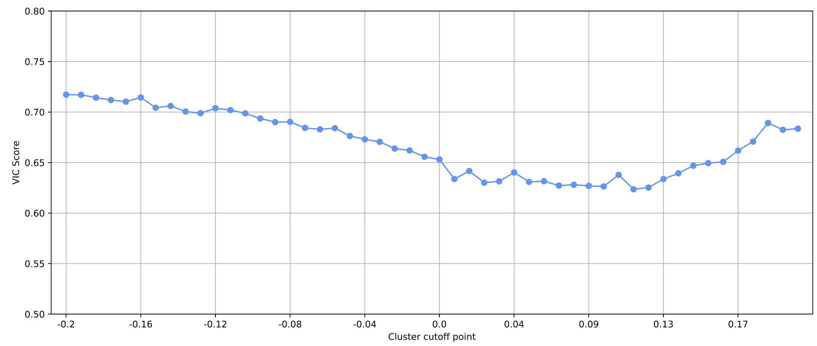

Fig. 1. VIC score obtained for each partition. The cluster cutoff point defines the division according to the score-change value.

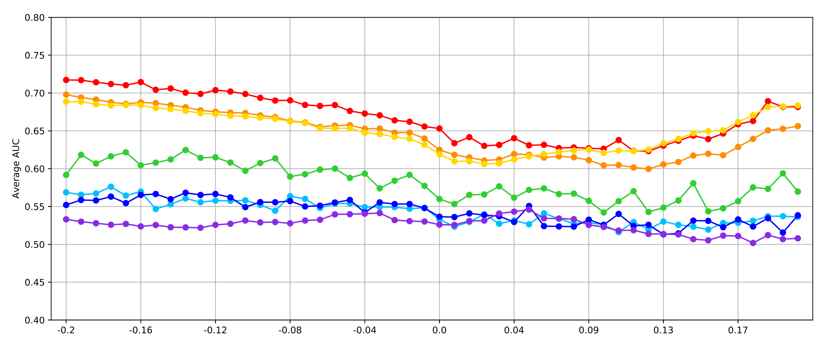

Fig. 2. Average AUC value for the 50 different binary partitions obtained by each supervised classifier. The average AUC was calculated with a 10 cross-validation. Random Forest: red; LDA: orange; Naïve Bayes: yellow; MLP: green; Decision Tree: blue; Extra Tree: indigo; k-NN: purple.

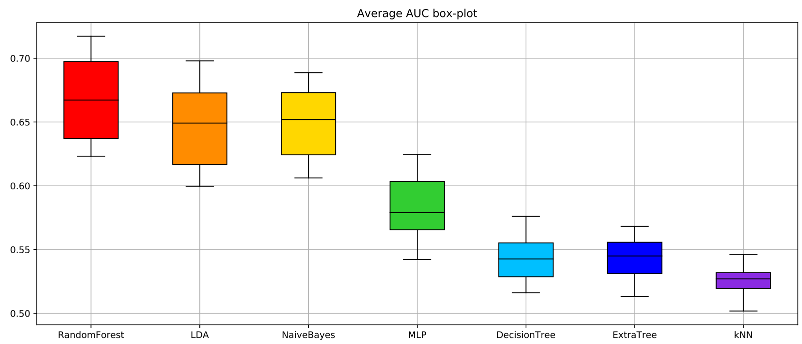

Fig. 3. Box plot of the average AUC for each classifier over all the binary partitions.

2.2 Tertiary Partition

The tertiary partitions were generated by selecting two cutoff points on the score-change attribute: one of them was a random floating point number over the interval $[-0.2, 0]$, and the other was randomly chosen over the interval $[0, 0.2]$. These numbers were verified so as to avoid cluster overlapping. In Table 2.2, the relevant information of the 50 tertiary partitions is presented. In addition, the VIC scores for each partition are shown in Fig. 4 and the AUC value for each classifier over the partitions are displayed in Fig. 5. Finally, the distribution of the AUC values of each classifier for all the partitions is shown in Fig. 6.

In contrast to the binary partitions, AUC values of the tertiary partitions only follow a close to constant trend independently of the supervised classifier used in the analysis (see Fig. 5). On the other hand, VIC scores are the same as the AUC of Random Forest, except for the partition 45. Hence, this suggests that the VIC score is independent of the cluster size. However, the best VIC scores were obtained when Class 1 had the biggest size, and Class 0 and Class 2 had similar sizes. On the contrary, the lowest VIC scores were obtained when Class 1 had the smallest size, and Class 0 and Class 2 were imbalanced, for instance, see partition number 36 and number 37 in Table 2.2. The highest VIC score was 0.679 for partitions number 10 and number 36 with cutoff points of -0.19 and 0.05, and -0.19 and 0.07, respectively. The cluster sizes were $\sim$1200 vs $\sim$2800 vs $\sim$1200 observations. Finally, as mentioned above, the best classifiers were Random Forest, LDA and Naïve Bayes, as seen in Fig. 6.

3 Conclusions

In this project, binary and tertiary partitions of a dataset were used to understand the effect of the cluster sizes. This effect was evaluated with the VIC method and leveraging seven supervised classifiers (Random Forest, Decision

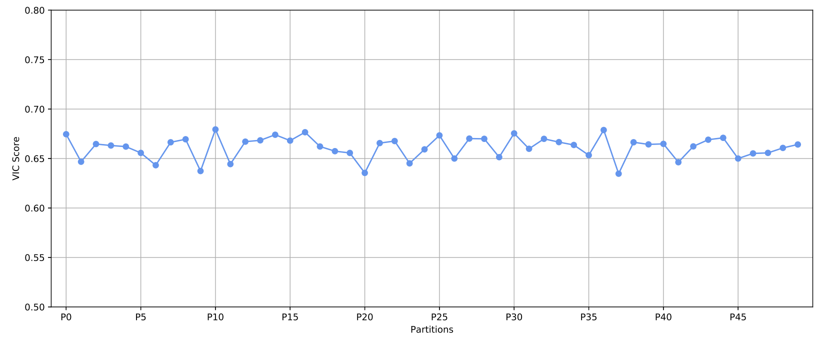

Fig. 4. VIC score obtained for each partition. The information of the partition is described in Table 2.2.

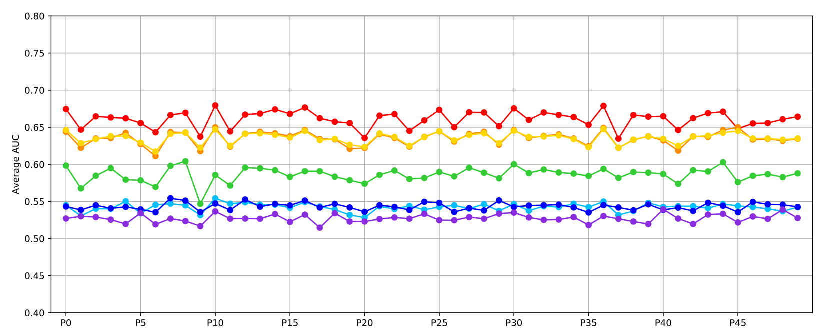

Fig. 5. Average AUC value for the 50 different tertiary partitions obtained by each supervised classifier. The average AUC was calculated with a 10 cross-validation. Random Forest: red; LDA: orange; Naïve Bayes: yellow; MLP: green; Decision Tree: blue; Extra Tree: indigo; *k*-NN: purple.

Table 2. Relevant information and results for the tertiary partitions.

Partition Number	Cutoff Value for 0-1 Class	Cutoff Value for 1-2 Class	Size of Class 0	Size of Class 1	Size of Class 2	VIC	Best Classifier
0	-0.17	0.08	1377	2829	1035	0.675	Random Forest
1	-0.05	0.03	2602	1086	1553	0.647	Random Forest
2	-0.1	0.04	2055	1743	1443	0.665	Random Forest
3	-0.07	0.08	2386	1820	1035	0.663	Random Forest
4	-0.18	0	1301	1949	1991	0.662	Random Forest
5	-0.03	0.04	2854	944	1443	0.656	Random Forest
6	-0.05	0	2602	648	1991	0.643	Random Forest
7	-0.11	0.07	1944	2181	1116	0.666	Random Forest
8	-0.15	0.09	1548	2730	963	0.669	Random Forest
9	-0.02	0	2965	285	1991	0.637	Random Forest
10	-0.19	0.05	1238	2689	1314	0.679	Random Forest
11	-0.01	0.04	3104	694	1443	0.644	Random Forest
12	-0.16	0.04	1446	2352	1443	0.667	Random Forest
13	-0.19	0.02	1238	2312	1691	0.668	Random Forest
14	-0.11	0.05	1944	1983	1314	0.674	Random Forest
15	-0.15	0.02	1548	2002	1691	0.668	Random Forest
16	-0.16	0.05	1446	2481	1314	0.677	Random Forest
17	-0.14	0	1641	1609	1991	0.662	Random Forest
18	-0.03	0.06	2854	1189	1198	0.657	Random Forest
19	-0.07	0.02	2386	1164	1691	0.656	Random Forest
20	0	0.09	3224	1054	963	0.635	Random Forest
21	-0.18	0.03	1301	2387	1553	0.666	Random Forest
22	-0.06	0.07	2496	1629	1116	0.668	Random Forest
23	-0.01	0.04	3104	694	1443	0.645	Random Forest
24	-0.03	0.07	2854	1271	1116	0.659	Random Forest
25	-0.14	0.06	1641	2402	1198	0.673	Random Forest
26	-0.02	0.08	2965	1241	1035	0.65	Random Forest
27	-0.1	0.06	2055	1988	1198	0.67	Random Forest
28	-0.19	0.02	1238	2312	1691	0.67	Random Forest
29	-0.04	0.04	2730	1068	1443	0.651	Random Forest
30	-0.17	0.05	1377	2550	1314	0.675	Random Forest
31	-0.06	0.07	2496	1629	1116	0.66	Random Forest
32	-0.13	0.04	1732	2066	1443	0.67	Random Forest
33	-0.1	0.08	2055	2151	1035	0.667	Random Forest
34	-0.1	0.04	2055	1743	1443	0.664	Random Forest
35	-0.1	0	2055	1195	1991	0.653	Random Forest
36	-0.19	0.07	1238	2887	1116	0.679	Random Forest
37	0	0.06	3224	819	1198	0.635	Random Forest
38	-0.12	0.02	1844	1706	1691	0.666	Random Forest
39	-0.1	0.1	2055	2285	901	0.664	Random Forest
40	-0.05	0.05	2602	1325	1314	0.665	Random Forest
41	-0.03	0.02	2854	696	1691	0.646	Random Forest
42	-0.1	0.1	2055	2285	901	0.662	Random Forest
43	-0.08	0.05	2270	1657	1314	0.669	Random Forest
44	-0.15	0.05	1548	2379	1314	0.671	Random Forest
45	-0.01	0	3104	146	1991	0.65	LDA
46	-0.04	0.08	2730	1476	1035	0.655	Random Forest
47	-0.02	0.07	2965	1160	1116	0.656	Random Forest
48	-0.04	0.05	2730	1197	1314	0.661	Random Forest
49	-0.09	0.04	2158	1640	1443	0.664	Random Forest

Tree, Extra Tree, Naïve Bayes, k-NN, MLP and LDA). The findings are that, in binary partitions, the VIC score and AUC values were higher when having extremely imbalanced cluster sizes. Additionally, in the tertiary partition, the cluster size did not have a vast effect over the VIC scores. However, the highest and the lowest scores depended on the size of Class 1 and a balanced size of

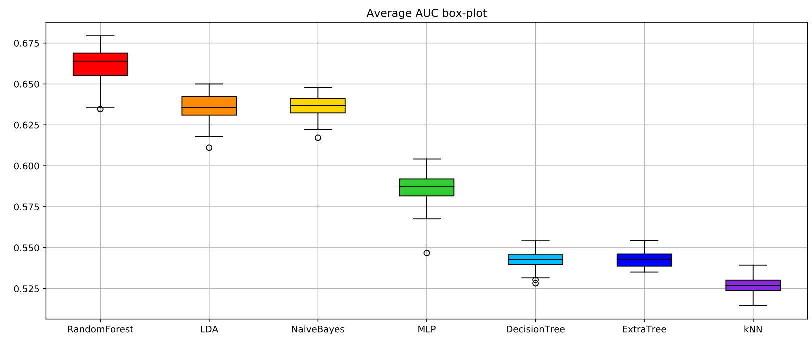

Fig. 6. Box plot of the average AUC for each classifier over all the tertiary partitions.

Class 0 and 2. In conclusion, the cluster size can or cannot have an effect on classification since it depends on the structure and the partition number of the dataset.

References

1. J. Rodríguez, M. A. Medina-Pérez, A. E. Gutierrez-Rodríguez, R. Monroy, and H. Terashima-Marín, "Cluster validation using an ensemble of supervised classifiers," *Knowledge-Based Systems*, vol. 145, pp. 134–144, 2018.
2. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
3. J. Huang and C. X. Ling, "Using AUC and accuracy in evaluating learning algorithms," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 17, no. 3, pp. 299–310, Mar. 2005.